PROYECTO INFORME FINAL IA

Integrantes:

Angela Maria Gómez Londoño CC: 1026159266
Manuela Zapata Cifuentes CC: 1040326871
Felipe Barrera Hincapié CC: 1007223451

1. Introducción

Al respirar una persona puede emitir sonidos que están directamente relacionados con el movimiento del aire dentro de la cavidad pulmonar, los cambios alrededor del tejido y las secreciones dentro del pulmón; Se conocen diferentes tipos de sonidos y la disminución o ausencia de estos puede significar aire o líquido alrededor de los pulmones, incremento del grosor de la pared torácica, demasiada insuflación o disminución del flujo de aire a una parte de los pulmones, algunos de los sonidos anormales presentes pueden ser sibilancias (frecuencias dominantes van alrededor de los 100 a 2000 Hz y sus rangos de duración comprenden desde los 80 hasta los 250 ms) o crepitancias (LS discontinuos, con una duración de al menos 20 ms y un rango de frecuencia típicamente desde 10 hasta 2000 Hz)

Es posible detectar estos sonidos usando estetoscopios digitales y diferentes técnicas de grabación, sin embargo los métodos digitales, tales como el reconocimiento de patrones y el procesamiento digital de señales han mostrado resultados más precisos, por lo cual se convierte en una valiosa herramienta para el diagnóstico automático de trastornos respiratorios como lo son por ejemplo, asma, neumonía y bronquiolitis.

Dado lo anterior, en el proyecto se desarrolló un modelo que permitiera relacionar de manera eficiente la presencia de sonidos respiratorios al cuantificar de manera eficiente el comportamiento y respuesta de estos.

Para ello, se utilizó el dataset llamado “Respiratory Sound DataBase” (enlace: https://www.kaggle.com/datasets/vbookshelf/respiratory-sound-database). El cual incluye sonidos respiratorios limpios y grabaciones ruidosas que se encargan de simular condiciones de la vida real.

Para el desarrollo del proyecto se utilizaron 3 métricas, precisión, recall y harmonic mean.

El accuracy da un valor entre 0 y 1, o en porcentaje entre 0 y 100%. Esta es la medida más directa de calidad entre las métricas, pero no siempre es la adecuada.

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive}$$

El recall o sensibilidad indica la proporción que el modelo de machine learning es capaz de identificar, en el caso del proyecto será qué cantidad de personas con sibilancias o crepitancias es capaz de identificar.

$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative}$$

F1 se calcula haciendo la media armónica entre la precisión y la exhaustividad, este se utiliza para combinar las medidas de precisión y recall en una sola métrica, esto se hace asumiendo que importa en igual medida precisión y el recall.

$$F1 = 2\frac{Precision * Recall}{Precision\ +\ Recall}$$

2. Exploración descriptiva del dataset

La exploración del dataset se realizó haciendo uso de un notebook de colab, el dataset a utilizar, incluye sonidos respiratorios limpios y grabaciones ruidosas que se encargan de simular condiciones de la vida real. En total hay 920 grabaciones, las cuales fueron tomadas de 126 pacientes que incluyen niños, adultos y ancianos, y tienen una duración variable de 10 a 90 segundos. Adicional a ello, de los 6898 ciclos respiratorios que se registraron durante 5.5 h, 1864 contienen estertores, 886 contienen sibilancias y 506 contienen crepitancias y sibilancias.

Este dataset incluye:

- 920 archivos de sonidos .wav
- Archivos de anotación 920.
- Archivo de texto que enumera el diagnóstico de cada paciente.
- Archivo de texto que enumera 91 nombres (filename_differences.txt).
- Archivo de texto que contiene información demográfica de cada paciente.

En el archivo “patient_diagnostic” se encontró que este se encuentra compuesto de 2 columnas. La primera es el ID de cada paciente, y la segunda es la enfermedad que tienen. En la columna de enfermedad se encuentran los posibles siguientes valores:

- COPD: Chronic Obstructive Pulmonary Disease
- LRTI: Lower Respiratory Tract Infection
- URTI: Upper Respiratory Tract Infection
- Asthma
- Healthy

	pid	Enfermedad
0	101	URTI
1	102	Healthy
2	103	Asthma
3	104	COPD
4	105	URTI

Figura 1. Contenido del archivo “patient_diagnostic”

Además fue posible conocer información del paciente relacionada con la edad, el género, y el índice de BMI(kg/m2), en la cual se clasifican los pacientes y las edades de estos, encontrándose como base la siguiente información respecto a los 125 pacientes:

- 36 bebés
- 7 infantes
- 7 adolescentes
- 1 adultos jóvenes
- 12 adultos
- 62 adultos mayores

Los 920 archivos txt según nos dice la descripción del dataset poseen 4 columnas que significan lo siguiente:

- Comienzo de los ciclos respiratorios
- Fin del ciclo respiratorio
- Presencia/ausencia de crepitancias (presencia=1, ausencia=0)
- Presencia/ausencia de sibilancias (presencia=1, ausencia=0)

3. Iteraciones de desarrollo

Preprocesado de datos:

Los nombres de los archivos nos brindan 5 informaciones diferentes separadas por un guión bajo. La información que nos brindan es: número del paciente, índice de grabación, ubicación, modo de adquisición y equipo de grabación.

De manera adicional, se creó otro data frame con el promedio de sibilancias y crepitancias en todos los audios para cada sujeto, y se unieron los datos anteriores con los obtenidos en un sólo data frame usando el método merge.

	inicio	fin	crepitancias	sibilancias	pid	modo	nombre	Enfermedad
0	1.862	5.718	0	1	160	mc	160_1b3_Al_mc_AKGC417L	COPD
1	5.718	9.725	1	1	160	mc	160_1b3_Al_mc_AKGC417L	COPD
2	9.725	13.614	0	1	160	mc	160_1b3_Al_mc_AKGC417L	COPD
3	13.614	17.671	0	1	160	mc	160_1b3_Al_mc_AKGC417L	COPD
4	17.671	19.541	0	0	160	mc	160_1b3_Al_mc_AKGC417L	COPD

Figura 3. dataframe final de archivos de texto.

Para el preprocesamiento de los audios se usaron los 920 archivos txt, los cuales, como ya se mencionó poseen 4 columnas, siendo en este caso las de mayor interés, el comienzo del ciclo respiratorio (“inicio”) y el fin del ciclo respiratorio (“fin”), con estos 2 datos se halló la longitud del ciclo al restar el tiempo en que inició y finalizó el ciclo, una vez se obtuvo este valor, se determinó la frecuencia respiratoria para posteriormente obtener un diagnóstico que permitiera clasificar la frecuencia como normal o alta, teniendo en cuenta que el rango normal está comprendido entre 12 y 25 ciclos/min. En el caso en que la frecuencia respiratoria estuviera dentro de los rangos normales se asignaba un 0, y en el caso contrario un 1, tal y como se muestra en la figura 4.

filename	inicio	fin	Longitud de ciclo	FR	Diagnostico FR	pid
101	9.002485	10.737364	1.734879	35.0	1	101
102	9.807231	11.322615	1.515385	40.0	1	102
103	8.778667	12.047833	3.269167	18.0	0	103
104	8.795221	11.078845	2.283624	26.0	1	104
105	9.198375	11.682250	2.483875	24.0	0	105
...	...	...	...	...	...	...
222	35.535680	40.488657	4.952977	12.0	0	222
223	15.320182	18.412637	3.092455	19.0	0	223
224	9.680714	12.529643	2.848929	21.0	0	224
225	9.587714	11.012214	1.424500	42.0	1	225
226	9.012006	10.942506	1.930500	31.0	1	226

Figura 4. dataframe ciclos respiratorios.

Para el preprocesamiento de archivos de datos demográficos, se generó un dataset con los datos contenidos en el archivo “demographic_info.txt” de los cuales se extrajeron datos como la edad, el género y el índice de masa corporal. Donde para el diagnóstico los pacientes sanos se identificaban por un 0 y los pacientes enfermos con un 1. Estos datos se ingresaron a un dataframe como se muestra a continuación.

Es importante resaltar que se eliminaron ciertas columnas como lo es el 'BMI (kg/m2)','Child Height (cm)','Peso (kg)', dado que no aportan información relevante al problema.

	pid	Edad	Genero	Enfermedad	Diagnostico
0	101	3.00	F	URTI	1
1	102	0.75	F	Healthy	0
2	103	70.00	F	Asthma	1
3	104	70.00	F	COPD	1
4	105	7.00	F	URTI	1
...	...	...	...	...	...
121	222	60.00	M	COPD	1
122	223	0.00	0	COPD	1
123	224	10.00	F	Healthy	0
124	225	0.83	M	Healthy	0
125	226	4.00	M	Pneumonia	1

Figura 5. dataframe datos demográficos.

Ahora bien, para la iteración 1 se utilizaron los datos demográficos presentes en la figura 5, donde en el caso del genero, se reemplazo el sexo femenino por un 0 y el masculino por 1, esto con el objetivo de tener sólo valores numericos como se muestra en la figura 6.

	pid	Edad	Genero	Enfermedad	Diagnostico
0	101	3.00	F	URTI	1
1	102	0.75	F	Healthy	0
2	103	70.00	F	Asthma	1
3	104	70.00	F	COPD	1
4	105	7.00	F	URTI	1
...	...	...	...	...	...
121	222	60.00	M	COPD	1
122	223	0.00	0	COPD	1
123	224	10.00	F	Healthy	0
124	225	0.83	M	Healthy	0
125	226	4.00	M	Pneumonia	1

Figura 6. Datos utilizados en la iteración 1.

Para la iteración 2, se añadieron 2 columnas extras que corresponden al promedio de las sibilancias y crepitantes que se escucharon en los audios ya mencionados (ver figura 7).

	pid	Crepitancias_mean	Sibilancias_mean	Edad	Genero	Enfermedad	Diagnostico
0	101	0.000000	0.000000	3.00	1	URTI	1
1	102	0.000000	0.000000	0.75	1	Healthy	0
2	103	0.000000	0.666667	70.00	1	Asthma	1
3	104	0.018182	0.181818	70.00	1	COPD	1
4	105	0.000000	0.000000	7.00	1	URTI	1
...	...	...	...	...	...	...	...
121	222	0.288889	0.177778	60.00	1	COPD	1
122	223	0.311475	0.213115	0.00	0	COPD	1
123	224	0.000000	0.000000	10.00	1	Healthy	0
124	225	0.000000	0.000000	0.83	1	Healthy	0
125	226	0.419355	0.000000	4.00	1	Pneumonia	1

Figura 7. Datos utilizados en la iteración 2.

Por último, para la iteración 3 se añadieron al data frame anterior 2 columnas correspondientes a la frecuencia respiratoria y al diagnóstico obtenido de la misma, como se muestra en la figura 8.

	pid	Crepitancias_mean	Sibilancias_mean	Edad	Genero	Enfermedad	Diagnostico	FR	Diagnostico FR
0	101	0.000000	0.000000	3.00	1	URTI	1	35	1
1	102	0.000000	0.000000	0.75	1	Healthy	0	40	1
2	103	0.000000	0.666667	70.00	1	Asthma	1	18	0
3	104	0.018182	0.181818	70.00	1	COPD	1	26	1
4	105	0.000000	0.000000	7.00	1	URTI	1	24	0
...	...	...	...	...	...	...	...	...	...
121	222	0.288889	0.177778	60.00	1	COPD	1	12	0
122	223	0.311475	0.213115	0.00	0	COPD	1	19	0
123	224	0.000000	0.000000	10.00	1	Healthy	0	21	0
124	225	0.000000	0.000000	0.83	1	Healthy	0	42	1
125	226	0.419355	0.000000	4.00	1	Pneumonia	1	31	1

Figura 8. Datos utilizados en la iteración 3

Iteración 1:

- **3.1.1. Modelo supervisado:** Se utilizó una red neuronal “MLPClassifier” con los siguientes parámetros,MLPClassifier(hidden_layer_sizes=(8,8,8), activation='relu', solver='adam', max_iter=500). Se destinó un 70% de los datos para train y un 30% para test.

Resultados, métricas y curvas de aprendizaje: Después de haber aplicado el modelo con los datos de test y entrenamiento que se seleccionaron, se obtuvieron los siguientes resultados.

	precision	recall	f1-score	support
Sanos	0.33	0.75	0.46	4
Enfermos	0.97	0.82	0.89	34
accuracy			0.82	38
macro avg	0.65	0.79	0.68	38
weighted avg	0.90	0.82	0.84	38

Figura 9. Métricas de desempeño iteración 1

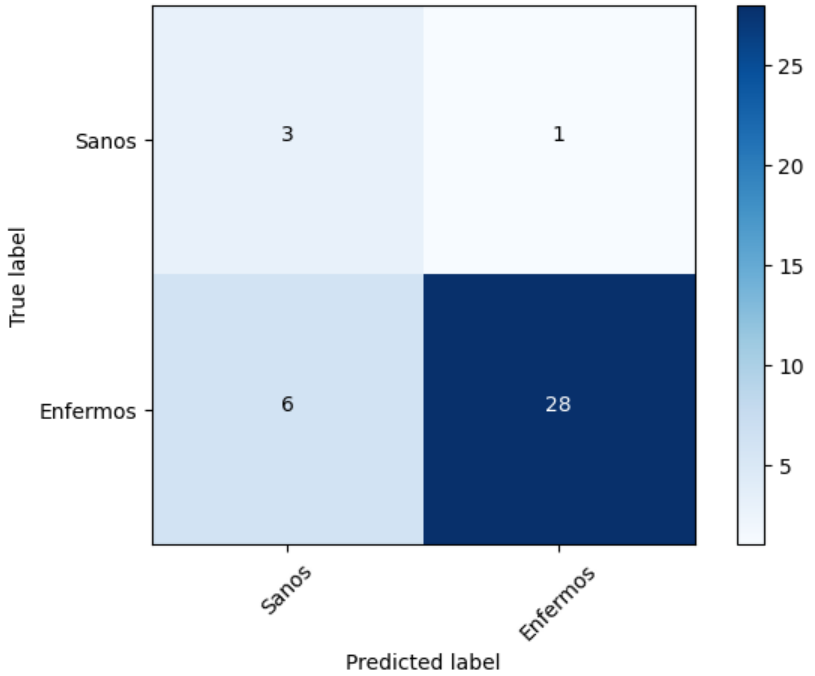

Figura 10. Matriz de confusión iteración 1

Para la clase "Sanos", el modelo tiene una precisión de 0.33, lo que significa que de todas las instancias clasificadas como "Sanos", solo el 33% son realmente "Sanos". Sin embargo, el recall (sensibilidad) es alto con un valor de 0.75, lo que indica que el modelo identifica correctamente el 75% de las instancias verdaderamente "Sanas". El puntaje F1 de 0.46 refleja un equilibrio entre precisión y recall para esta clase.

Para la clase "Enfermos", el modelo tiene una precisión alta de 0.97, lo que significa que de todas las instancias clasificadas como "Enfermos", el 97% son realmente "Enfermos". El recall es de 0.82, lo que indica que el modelo identifica correctamente el 82% de las instancias verdaderamente "Enfermas". El puntaje F1 de 0.89 muestra un buen equilibrio entre precisión y recall para esta clase.

El accuracy (exactitud) general del modelo es de 0.82, lo que indica que el modelo clasifica correctamente el 82% de las instancias en general.

En resumen, el modelo parece tener un buen rendimiento en la clasificación de la clase "Enfermos", con una alta precisión y recall. Sin embargo, su rendimiento en la clasificación de la clase "Sanos" es inferior, con una baja precisión pero un recall aceptable.

Iteración 2:

- **3.1.2. Modelo supervisado:** Se utilizó una red neuronal “MLPClassifier” con los siguientes parámetros,MLPClassifier(hidden_layer_sizes=(8,8,8), activation='relu', solver='adam', max_iter=500). Se destinó un 70% de los datos para train y un 30% para test.

  Resultados, métricas y curvas de aprendizaje: Después de haber aplicado el modelo con los datos de test y entrenamiento que se seleccionaron, se obtuvieron los siguientes resultados.

	precision	recall	f1-score	support
Sanos	0.33	0.50	0.40	4
Enfermos	0.94	0.88	0.91	34
accuracy			0.84	38
macro avg	0.64	0.69	0.65	38
weighted avg	0.87	0.84	0.86	38

Figura 11. Métricas de desempeño iteración 2.

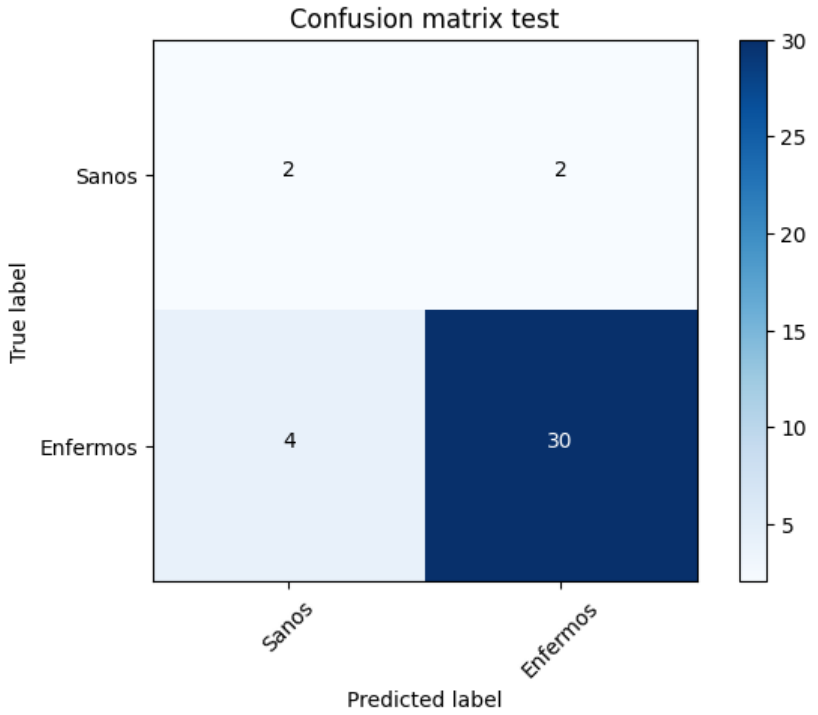

Figura 12. Matriz de confusión iteración 2.

De los resultados obtenidos para la iteración 2, se tiene que el modelo presenta una precisión del 33% para la clase "Sanos", lo que indica que todas las instancias clasificadas como sanos, esto es 2, en realidad lo están, tal y como se puede observar en la matriz de confusión presente en la figura 12. Para la métrica recall y valor-F1 se obtuvieron resultados, 50% y 40% respectivamente, lo que da cuenta de que el modelo no discrimina de manera precisa a los pacientes sanos de los pacientes enfermos.

Para la clase "Enfermos" se obtuvo una precisión, un recall, y un valor-F1 igual a 94%, 88%, 91%, respectivamente, lo que indica que los pacientes clasificados como enfermos, esto es 30, en realidad lo están. De manera adicional, al tener un recall del 88% muy pocos pacientes enfermos fueron clasificados como sanos, esto puede corroborarse en la matriz de confusión. Este resultado nos indica que este modelo es bueno identificando pacientes enfermos pero no sanos.

Para esta interacción se consiguió una exactitud del 84%, pero esto no indica que el modelo sea bueno, ya que no obtuvo buenos resultados en la clasificación. Es por esto que es necesario cambiar el modelo para intentar obtener mejores resultados.

Iteración 3:

- **3.1.3. Modelo supervisado:** Se utilizó una red neuronal “MLPClassifier” con los siguientes parámetros,MLPClassifier(hidden_layer_sizes=(8,8,8), activation='relu',

solver='adam', max_iter=500). Se destinó un 70% de los datos para train y un 30% para test.

Resultados, métricas y curvas de aprendizaje: Después de haber aplicado el modelo con los datos de test y entrenamiento que se seleccionaron, se obtuvieron los siguientes resultados.

	precision	recall	f1-score	support
Sanos	0.96	0.92	0.94	25
Enfermos	0.86	0.92	0.89	13
accuracy			0.92	38
macro avg	0.91	0.92	0.91	38
weighted avg	0.92	0.92	0.92	38

Figura 13. Métricas de desempeño iteración 3.

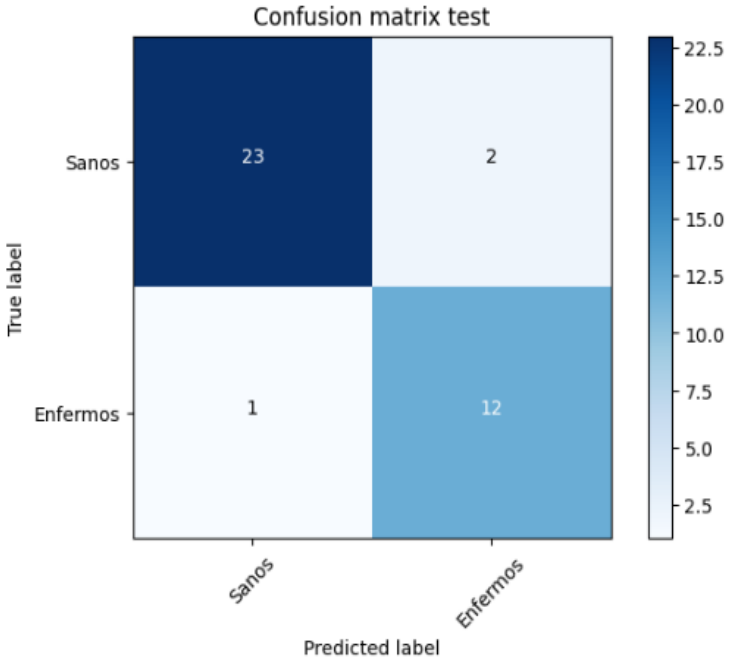

Figura 14. Matriz de confusión iteración 3.

De los resultados obtenidos para la iteración 3, se tiene que el modelo presenta una precisión del 96% para la clase "Sanos", lo que indica que todas las instancias clasificadas como sanos, esto es 23, en realidad lo están, tal y como se puede observar en la matriz de confusión presente en la figura 14. De igual manera, para la métrica recall y valor-F1 se obtuvieron buenos resultados, 92% y 94%, lo que da cuenta de que el modelo discrimina de manera precisa a los pacientes sanos de los pacientes enfermos, debido a que al tener un recall del 92% muy pocos pacientes sanos fueron clasificados como enfermos.

Para la clase "Enfermos" se obtuvo una precisión, un recall, y un valor-F1 igual a 86%, 92%, 89%, respectivamente, lo que indica que los pacientes clasificados como enfermos, esto es 12, en realidad lo están. De manera adicional, al tener un recall del 92% muy pocos pacientes enfermos fueron clasificados como sanos, esto puede corroborarse en la matriz de confusión.

Para esta interacción se consiguió una exactitud del 92%, lo que da cuenta de que los resultados obtenidos estuvieron cerca a los valores verdaderos. Por lo que de las 3 iteraciones realizadas, este modelo es el que mejores resultados ofrece, pues para ambas clases, sanos y enfermos se presentaron valores de precisión mayores al 80%, y valores de recall y valor-F1 altos, por lo que se asemejan a los verdaderos.

4. Retos y consideraciones de despliegue

Los modelos MLP pueden volverse complejos y computacionalmente intensivos, especialmente si se utilizan varias capas ocultas y un gran número de neuronas. Esto puede suponer un desafío en términos de tiempo de entrenamiento y recursos computacionales necesarios para entrenar y utilizar el modelo, de igual manera las herramientas necesarias para el machine learning supusieron un reto debido a la gran cantidad de conocimientos que requieren

Algunas de las consideraciones de despliegue que se tienen son los siguientes:

- Para implementar el modelo en un entorno de producción en el que se requiera clasificar datos en tiempo real, es importante considerar la escalabilidad y el rendimiento del modelo. Esto implica garantizar que el modelo pueda manejar de manera eficiente un alto volumen de solicitudes y proporcionar predicciones rápidas sin afectar la latencia del sistema.
- Las consideraciones éticas puesto que se utilizara en el campo de la salud o cualquier otro ámbito sensible, y es importante tener en cuenta las consideraciones éticas y de privacidad. Esto puede implicar garantizar la protección de los datos sensibles, el cumplimiento de las regulaciones de privacidad y la transparencia en el uso y los resultados del modelo.
- Una vez desplegado, es esencial establecer un mecanismo de monitoreo para evaluar el rendimiento y la precisión del modelo en producción. También puede ser necesario implementar un proceso de actualización del modelo para incorporar nuevos datos y mejorar continuamente la precisión y el rendimiento del sistema.

5. Conclusiones

- El potencial del machine learning y la inteligencia artificial en proyectos futuros es amplio y diverso. Estas tecnologías permiten realizar desde predicciones en el mercado financiero hasta diagnósticos médicos. Su versatilidad se debe a la variedad de enfoques y parámetros que se pueden ajustar para obtener modelos adaptados a cada caso. Además, la accesibilidad creciente a través de internet y la disponibilidad de cursos gratuitos facilitan su aprendizaje y aplicación para un público más amplio.
- Es fundamental contar con conjuntos de datos relevantes y bien organizados para aprovechar estas herramientas. La calidad y la utilidad de los datos son cruciales para entrenar modelos precisos. Además, tener conocimiento del dominio o contar con el apoyo de expertos en el proyecto ayuda a organizar los datos de manera efectiva. No siempre es necesario utilizar un gran número de parámetros, ya que muchos de ellos pueden ser irrelevantes y generar cambios innecesarios en el modelo.
- La correcta preparación y gestión de los datos son aspectos fundamentales para el éxito de los proyectos de machine learning y inteligencia artificial. Sin datos organizados y significativos, incluso la mejor tecnología carecerá de utilidad. Es esencial contar con conjuntos de datos relevantes, limpios y correctamente estructurados para entrenar modelos precisos y obtener resultados confiables. La comprensión del dominio y la colaboración con expertos en el campo son elementos clave para garantizar una organización óptima de los datos y una selección adecuada de las características más relevantes. Así, se maximiza el potencial de estas poderosas herramientas en la toma de decisiones y la solución de problemas en diversos campos de aplicación.